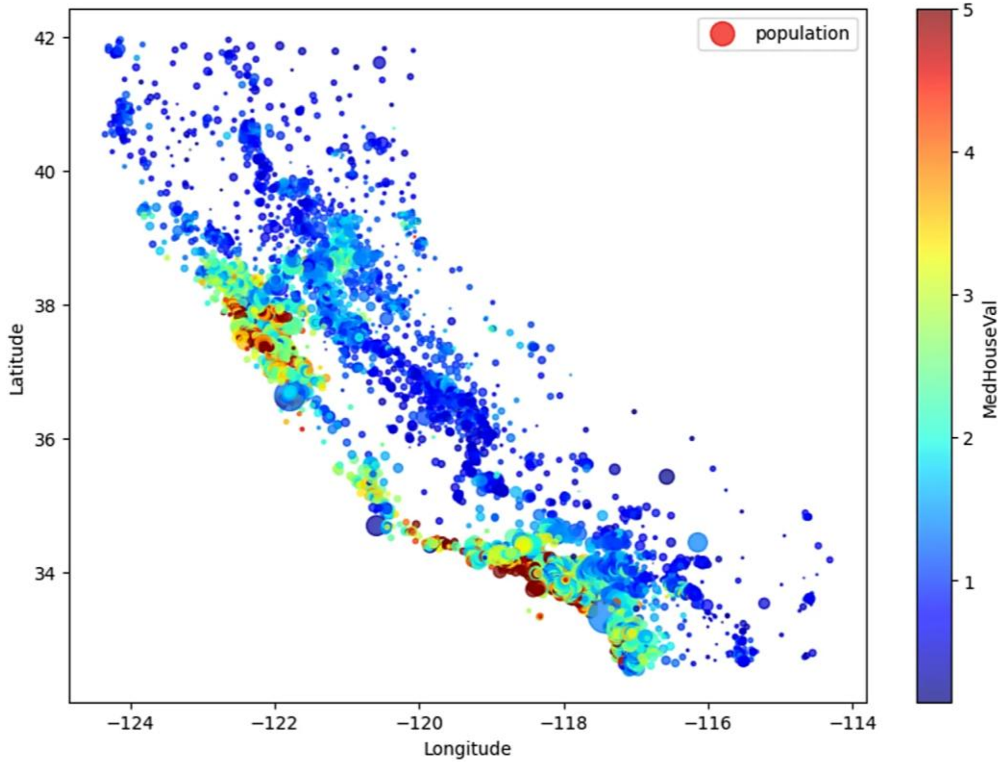
End-to-end Project

Machine Learning Course

Saeed Mohagheghi

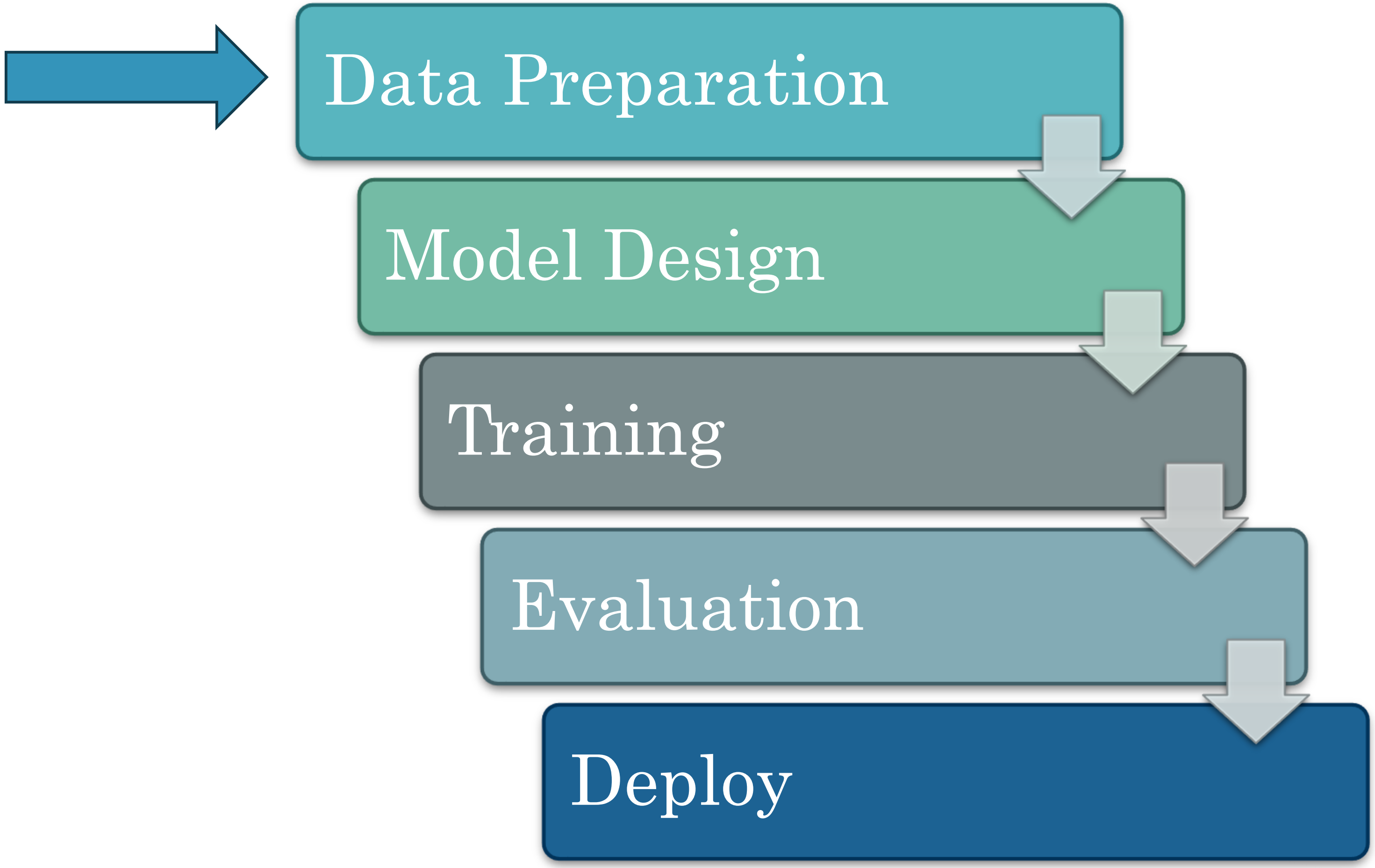
Project

- Predict **House Price** in California (in 1990!)
- Dataset: **California Housing** [download link]
  - Source: 1990 U.S. Census data
  - Instances: 20,640 districts
  - **Features**: 8 numeric / 1 categorical:
    - Latitude, Longitude
    - Median housing age
    - Total rooms, Total bedrooms
    - Population
    - Households
    - Median income
    - Ocean proximity (categorical)
  - **Target**: Median house value



→ Book's Chapter 2

ML Pipeline



Data Preparation

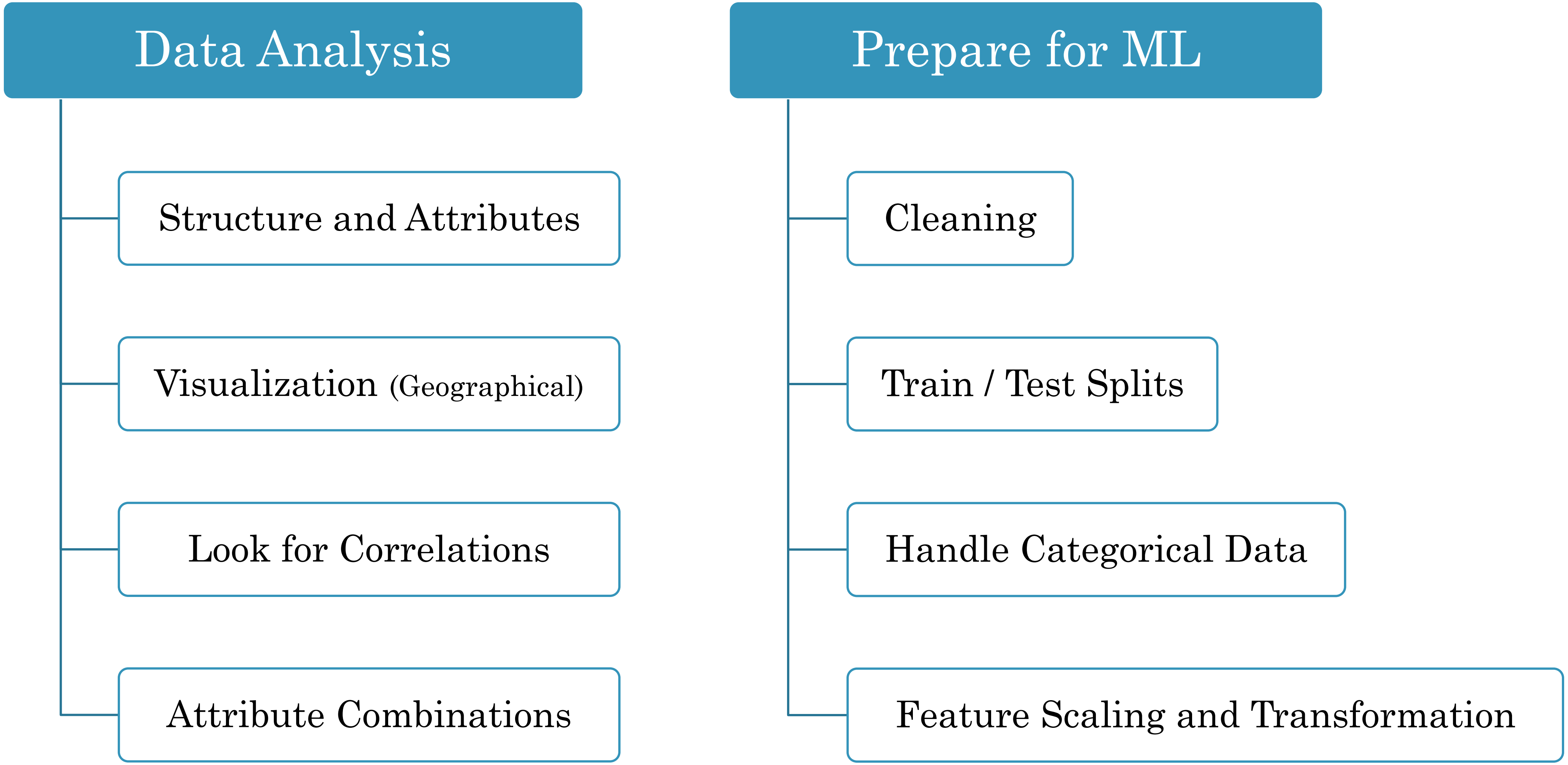
Model Design

Training

Evaluation

Deploy

Data Preparation



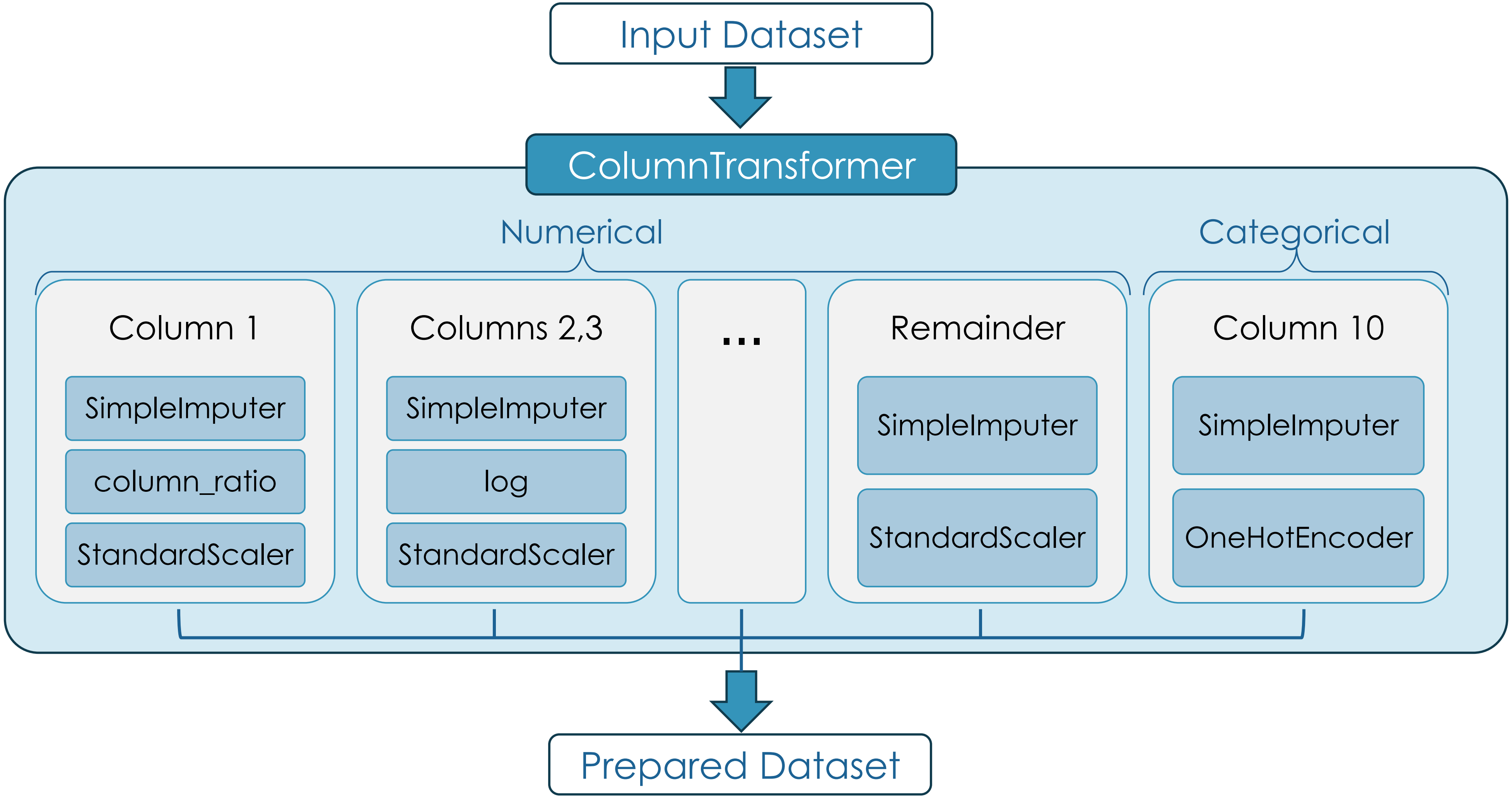
Data Analysis

- Structure and Attributes
- Visualization (Geographical)
- Look for Correlations
- Attribute Combinations

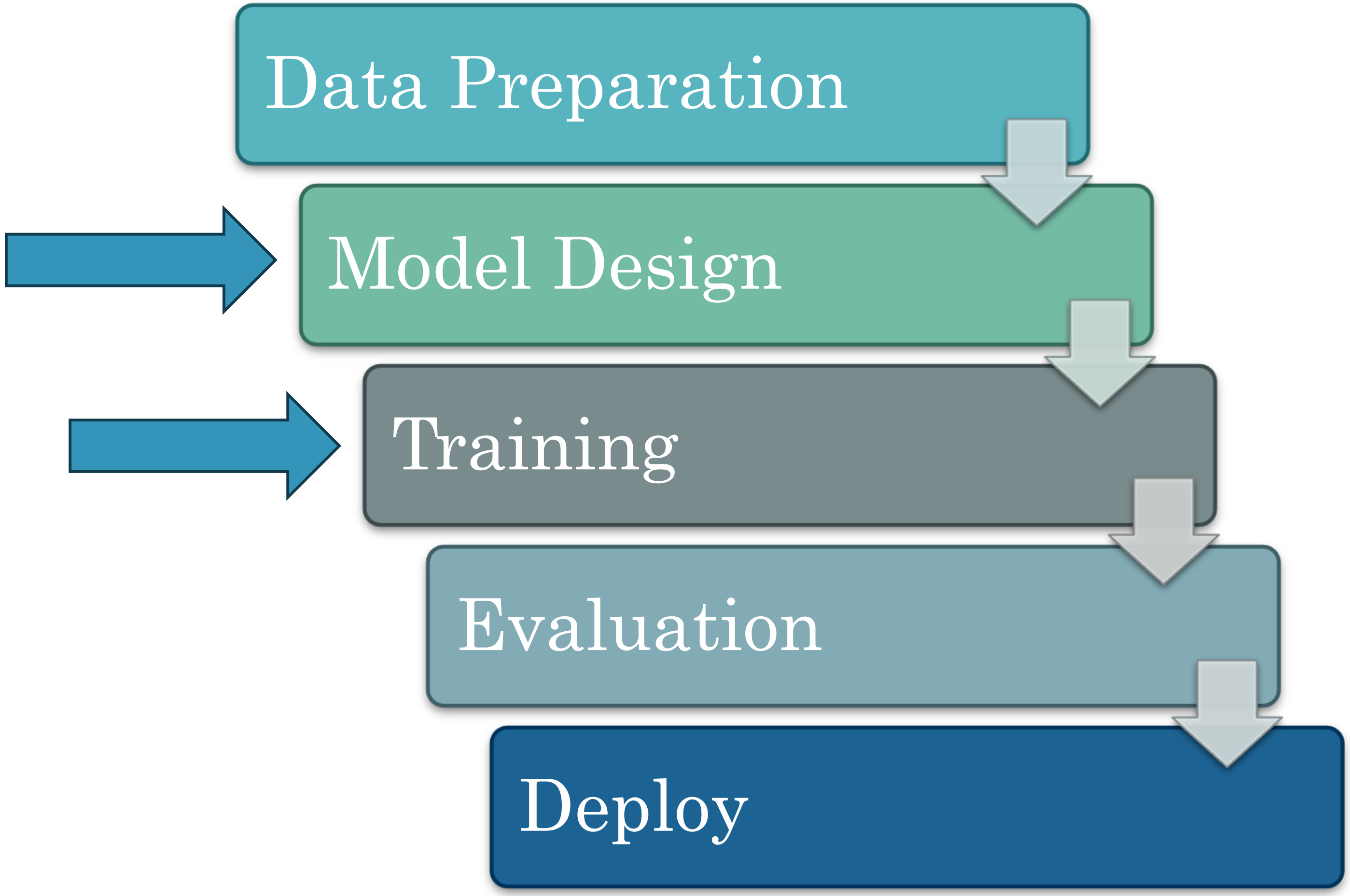
Prepare for ML

- Cleaning
- Train / Test Splits
- Handle Categorical Data
- Feature Scaling and Transformation

Data Preparation Pipeline



ML Pipeline



Data Preparation

Model Design

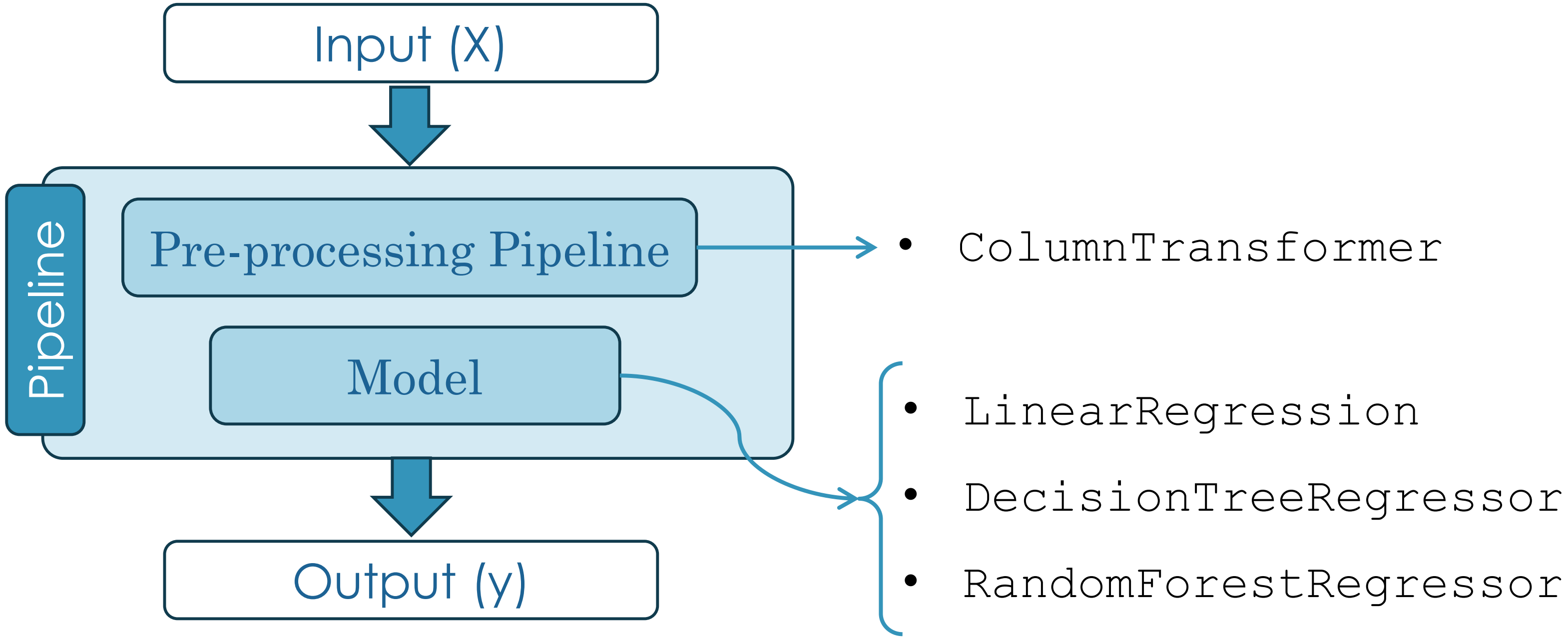
Training

Evaluation

Deploy

Full Pipeline

- Full Pipeline
  - to be used in Training / Evaluation / Deployment



Input (X)

Pipeline

Pre-processing Pipeline

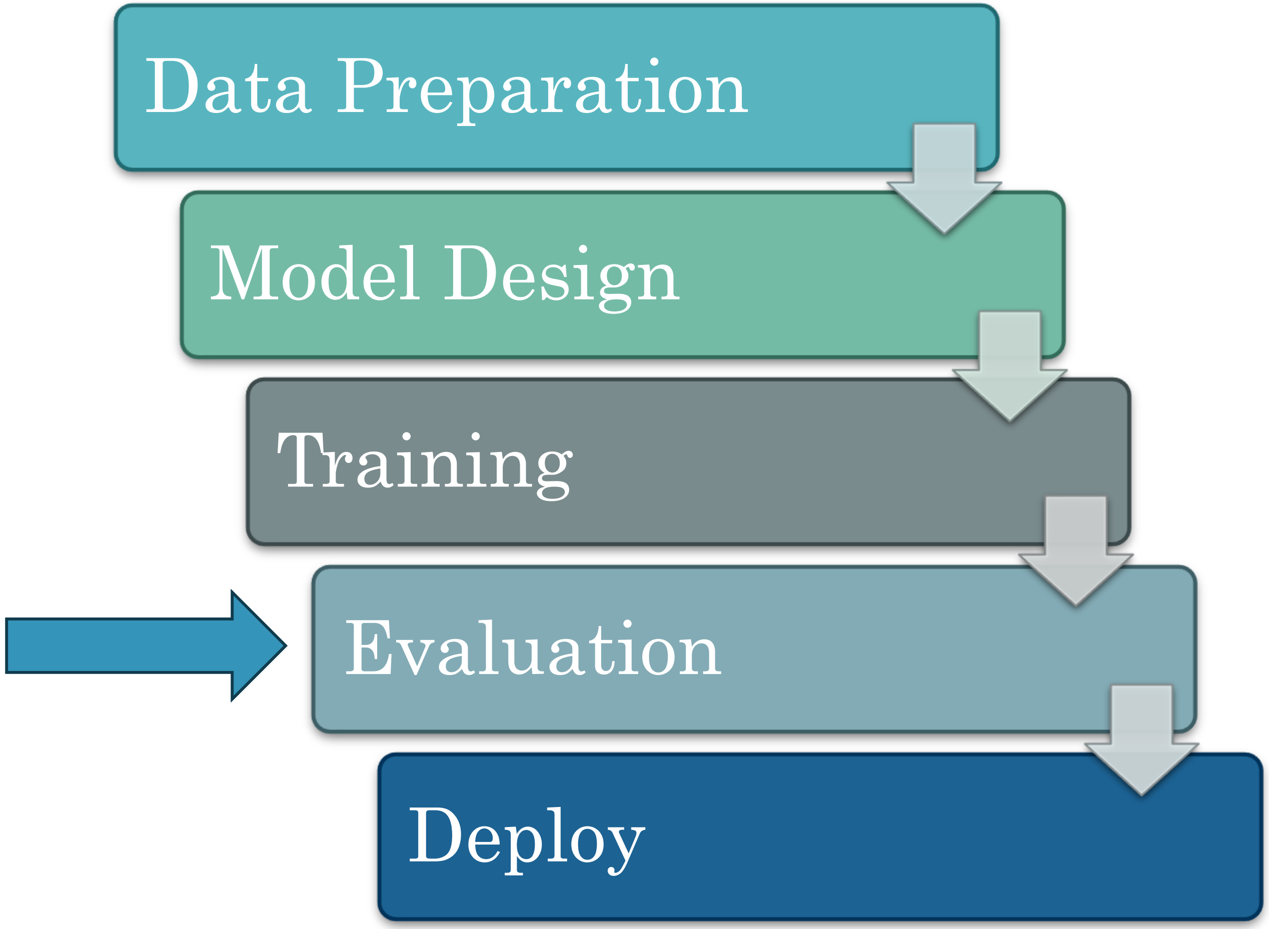
Model

Output (y)

- ColumnTransformer

- LinearRegression
- DecisionTreeRegressor
- RandomForestRegressor

ML Pipeline



Data Preparation

Model Design

Training

Evaluation

Deploy

Evaluation

- Using Metrics (Evaluate on Train/Val/Test Subset Separately)

    -
      ```
      from sklearn.metrics import root_mean_squared_error
      predicted_labels = pipeline.predict(housing)
      rmse = root_mean_squared_error( true_labels, predicted_labels )
      ```

- Using K-fold Cross-validation (Multiple Evaluations on Nonoverlapping Subsets)

    -
      ```
      from sklearn.model_selection import cross_val_score
      rmses = -cross_val_score( pipeline, input_data, true_labels,
                                scoring="neg_root_mean_squared_error",
                                cv=k )
      ```